**МИНОБРНАУКИ РОССИИ**

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики - процессов управления**

**Программа бакалавриата**

**“Большие данные и распределенная цифровая платформа”**

**ОТЧЕТ**

**по лабораторной работе №2**

**по дисциплине «Алгоритмы и структуры данных»**

**на тему «Обезличивание синтетических данных»**

**Вариант – 2**

**Студент гр. 23Б15-пу**

**Кубякин Н.А.**

**Преподаватель**

**Дик А.Г.**

**Санкт-Петербург**

**2024 г.**

***Оглавление***

[1. Цель работы 3](#_Toc178797423)

[2. Описание задачи (формализация задачи) 3](#_Toc178797424)

[3. Теоретическая часть 4](#_Toc178797431)

[4. Основные шаги программы 7](#_Toc178797432)

[5. Блок-схема программы 8](#_Toc178797433)

[6. Описание программы 11](#_Toc178797442)

[7. Рекомендации пользователя 12](#_Toc178797443)

[8. Рекомендации программиста 12](#_Toc178797444)

[9. Исходный код программы 13](#_Toc178797445)

[10. Контрольный пример 13](#_Toc178797447)

[11. Вывод 15](#_Toc178797448)

[12. Источники 15](#_Toc178797450)

# Цель работы

# Цель работы состоит в разработке программного решения для обезличивания датасета с учетом введенных квази-идентификаторов и реализации алгоритма k-анонимизации. Это предполагает создание программы с графическим интерфейсом, который обеспечит удобство использования для пользователя.

# Описание задачи (формализация задачи)

# Задача состоит в обезличивании данного датасета с учётом выбранных квази – идентификаторов и выполнении алгоритма k – анонимизации учитывая следующие условия:

# Программа должна считывать исходный файл (итоговый файл из 1-ой лабораторной работы).

# Программа делится по функционалу на обезличивание и вычисление k – anonymity.

# Пользователь может указывать квази – идентификаторы.

# Обезличивание проходит по нескольким методам из списка: локальное обобщение, агрегация, возмущение, микро-агрегация, перемешивание, создание псевдонимов, маскеризация, локальное подавление, удаление атрибутов, метод декомпозиции.

1. Используя метод k – anonymity рассчитать K для обезличенного набора.
2. Вывести 5 (или менее если их меньше 5) плохих значений К и их значения в процентах.
3. Вывести уникальные (К = 1) строки и их количество.
4. Вывести приемлемое К: >= 10 для файлов до 51000 строк, >= 7 для файлов до 105000 строк, >= 5 для файлов до 260000 строк.
5. Оценить полезность данных путём сравнения обезличенного набора с исходным

# Теоретическая часть

K-анонимность - это свойство релиза данных, при котором для каждой группы значений атрибутов, называемой квази-идентификатором, в наборе данных гарантируется, что каждая комбинация значений квази-идентификаторов встречается не менее k раз, где k - заданный параметр безопасности. Таким образом, это обеспечивает защиту конфиденциальности, предотвращая возможность однозначной идентификации отдельных субъектов данных. Более подробно тема k-анонимности расписана в статье [[1]](#stat).

В контексте датасета поликлиники, который является результатом лабораторной работы №1, необходимо рассмотреть существующие методы обезличивания данных и выбрать подходящие для этого конкретного набора данных.

* 1. Локальное обобщение – В рамках данной техники предполагается уменьшение специфичности атрибута за счет подмены точного значения атрибута его общим значением .
  2. Агрегация – Техника основана на построении обобщений по одному или нескольким атрибутам (групповые операции). Представляет из себя построение обобщенных данных по выделенным атрибутам.
  3. Возмущение – внесение шума в данные, которые перестают быть точными или правдивыми, но сохраняют основные статистические закономерности. Метод применим для бинарных данных (да/нет, например, пола), статистически частых наборов данных. Позволяет сохранить статистическую ценность набора данных при незначительных потерях информации.
  4. Микро-агрегация – Техника предполагает формирование групп записей, для которых вместо точных значений указывается диапазон. Метод рекомендуется к применению для параметров возраста, географических локаций, финансовых транзакций, биометрических данных.
  5. Перемешивание – Техника направлена на перемешивание аналогичных записей при сохранении статической значимости всего набора. Наиболее полезна при обезличивании временных интервалов, например, информации об отпусках. Метод применяется, как правило, в совокупности с другими методами обезличивания.
  6. Создание псевдонимов – Для чувствительного атрибута проводится замена идентификаторов на условные обозначение или цифровые ключи. Как правило, проводится для частных данных. Используется в совокупности с другими методами обезличивания.
  7. Маскеризация – В рамках данной техники проводится замена части записи заполнителями по определенному шаблону. Хотя в результате исходной строке сопоставляется строка определенного шаблона, что позволяет отнести этот метод к псевдонимизации, по ряду характеристик этот метод может быть сопоставлен микро-агрегации: в результате применения шаблонов возникают обобщенные группы. Применим для идентификаторов (номера паспортов, телефоны, номера карт)
  8. Локальное подавление – удаление или перекодирование относительно редких записей данных. В отличии от метода удаления атрибутов нацелен на удаление строк или значений для заданных атрибутов в выделенных строках. Наибольшее применение имеет для медицинской и биометрической информации, а также больших наборов статистических данных с аномалиями.
  9. Удаление атрибутов – удаление чувствительного контента без добавления замен. Метод используется для удаления прямых идентификаторов, а также удаления избыточных квази-идентификаторов.
  10. Метод декомпозиции – Разделение множества атрибутов на несколько подмножеств с последующим раздельным хранением. Метод не является собственно техникой обезличивания, но позволяет формировать наборы данных, к которым могут применяться методы обезличивания.

Для каждого атрибута набора данных можно применить определенный метод обезличивания:

1. Атрибуты "Name" содержат информацию, позволяющую однозначно идентифицировать человека, поэтому для обезличивания данных оптимально использовать метод удаления атрибутов.

2. "Snils" содержат информацию, позволяющую однозначно идентифицировать человека, поэтому для обезличивания данных оптимально использовать метод удаления атрибутов.

3. "Passport": Данный атрибут также содержит информацию, по которой можно однозначно идентифицировать человека. Однако первые две цифры кода региона не относятся к конкретному человеку. Можно применить метод маскеризации, чтобы скрыть оставшуюся информацию.

4. "Symptoms" и "Analysis": Эти атрибуты однозначно связаны с атрибутами “Doctor”, поэтому для обезличивания данных оптимально использовать метод удаления атрибутов.

5. "Doctor": Информация о докторе может быть обобщена до отделения, в котором он работает.

6. "DateStart" и "DateEnd": Дату приема можно обобщить до времени года, а дату получения результата удалить.

7. "Price": Для атрибута стоимости услуг можно применить микро-агрегацию путем округления вниз до ближайшего кратного 2500.

8. "Card": Данные о банковской карте могут быть обобщены, учитывая, что часть цифр относится к платежной системе и банку.

# Основные шаги программы

1. С помощью интерфейса пользователь указывает квази – идентификаторы и вводит название файла с датасетом и запускает программу.
2. Датасет обезличивается.
3. Считается k – анонимность датасета.
4. Поиск 5-ти наименьших значений K и вычисление процентного количества строк с этими коэффициентами.
5. Поиск уникальных строк (с коэффициентом K = 1).
6. Вывод полученных результатов в интерфейс программы.

# Блок-схема программы

# На рисунках 1-3 представлена блок-схема алгоритма.

# 

# Рис. 5.1 Блок-схема основной функции.

# 

# Рис. 5.2 Блок-схема обезличивания датасета.

# 

# Рис. 5.3 Блок-схема метода вычисления k – анонимности.

# Описание программы

Программа реализована на языке Python 3.10 с использованием библиотек pandas [[2]](#pandas), tkinter [[3]](#tkinter).

В программе используются 7 функций для работы с данными и 1 функция для ввода исходного датасета и вывода результата.

Таблица 6.1 Описание функций

|  |  |  |
| --- | --- | --- |
|  | Имя функции | Назначение |
|  | maskerization\_passport | Маскеризация паспорта (XX\*\* \*\*\*\*\*\*). |
|  | generalization\_doctor | Обобщает врача до отделения, в котором он принимает. |
|  | generalization\_date\_start | Обобщает дату до времени года. |
|  | aggregation\_price | Производит агрегацию цены до интервалов в 2500 рублей. |
|  | generalization\_card | Обобщает карту до платёжной системы, к которой привязана карта. |
|  | anonimyze | Обезличивание датасета. |
|  | calc\_k\_anonimity | Вычисление k-анонимности. |
|  | run\_function | Получение вводной информации и вывод результата в интерфейс и создание файлов с обработанным датасетом. |

# Рекомендации пользователя

После запуска программы требуется:

1. Ввести название исходного датасета включая расширение в поле “Input”.
2. Выбрать хотя бы один квази – идентификатор проставив галочки рядом с выбранными полями.
3. Запустить программу с помощью кнопки “Run”.
4. Результаты будут выведены в поле внизу интерфейса, а в папку с программой будут сохранены файлы с уникальными строками (excess.csv), с обезличенным датасетом (full\_anon.csv), со списком всех строк с количеством их повторов (types.csv) и с датасетом с удаленными уникальными строками (full\_without\_excess.csv).

# Рекомендации программиста

Для запуска программы необходима 64-битная операционная система Windows, Linux или macOS. Для работы с кодом необходима среда разработки, совместимая с python 3.1 и библиотеки pandas [[2]](#pandas), tkinter [[3]](#tkinter).

Для запуска программы необходимо установить Python версии не ниже 3.10, а так же библиотеки pandas [[2]](#pandas), tkinter [[3]](#tkinter).

В папке с файлами с кодом должны располагаться исходные датасеты.

# Исходный код программы

# Исходный код программы и необходимые текстовые файлы доступны по ссылке: <https://github.com/NikiTaku1/spbu_alg/tree/main/lab2>

# Контрольный пример

В данном разделе представлен пример, демонстрирующий работу программы.

1) Запуск программы, открытие интерфейса  
2) Ввод названия файла (здесь: data250k.csv, файл на 250000 строк)

3) Выбор квази – идентификаторов (здесь: Passport, Doctor, DateStart, Price, Card)

4) Запуск программы кнопкой “Run”

5) Вывод результата в интерфейсе и сохранение файлов с информацией в папке с кодом

Работа программы и результаты представлены на рисунках 10.1 – 10.6

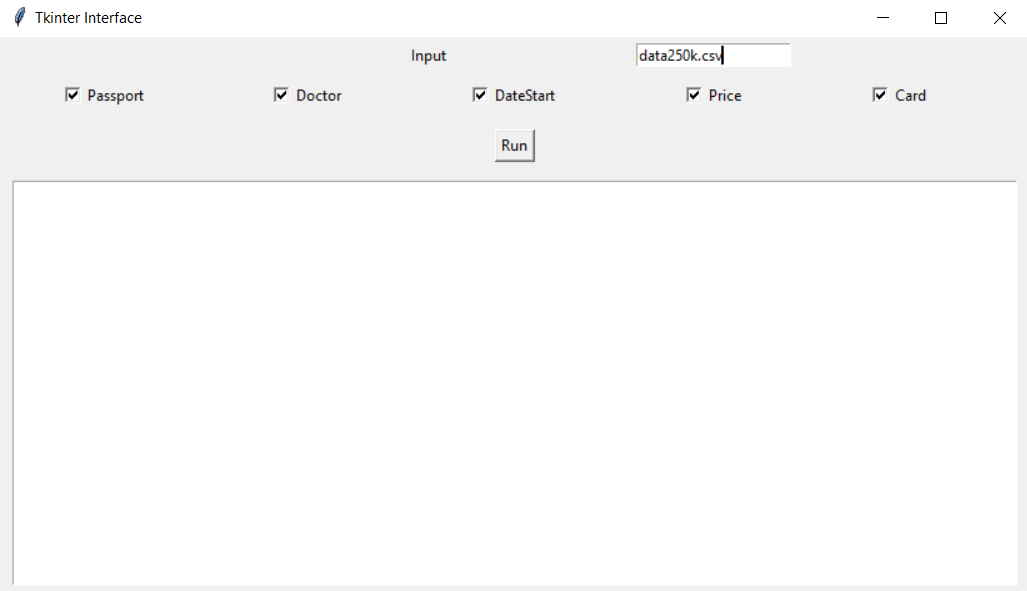


Рис. 10.1 Ввод данных в интерфейс программы

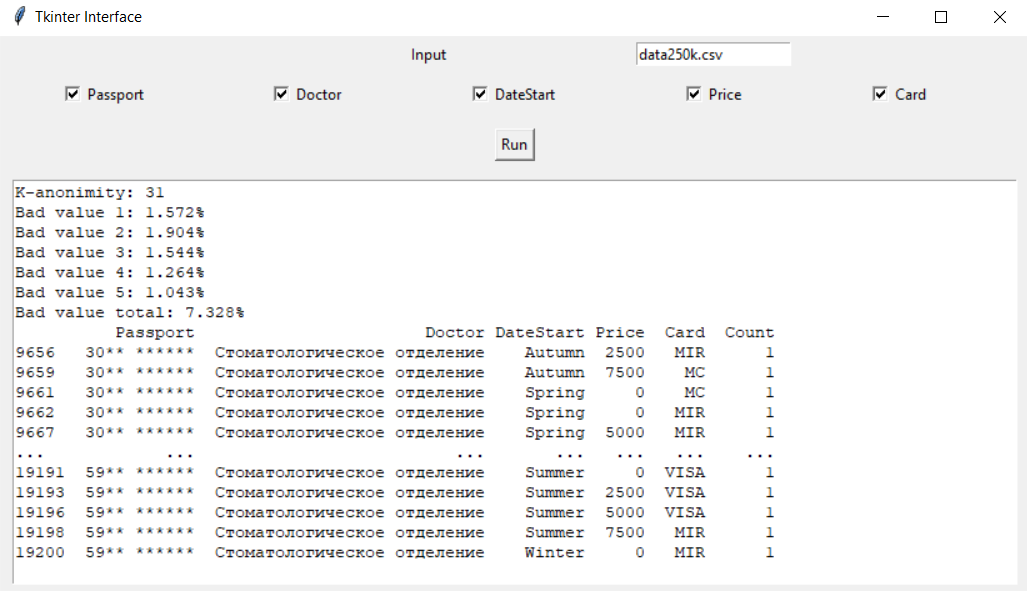


Рис. 10.2 Вывод результата в интерфейсе

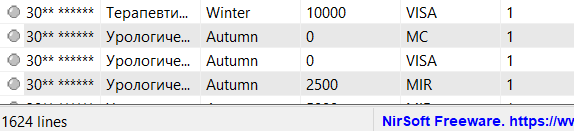


Рис 10.3 Файл с уникальными строками

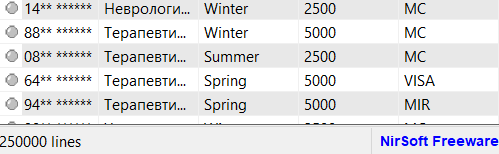


Рис 10.4 Файл с обезличенным датасетом

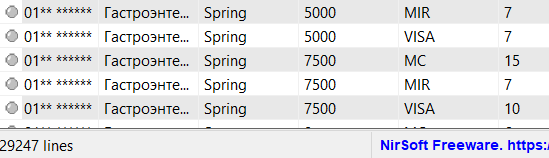


Рис 10.5 Файл с типами и количеством их повторов

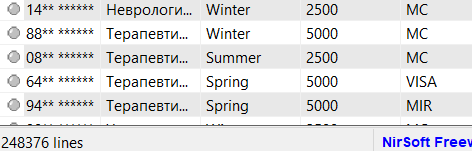


Рис 10.6 Исходный датасет без уникальных строк

# Вывод

# В результате выполнения работы изучена теоретическая информация о К – анонимности и методах обезличивания информации, разработана и написана программа, позволяющая с помощью интерфейса пользователя выбирать квази-идентификаторы, по которым происходит обезличивание таблицы, и выводится коэффициент анонимности, наименьшие значения К и уникальные строки.

# Источники

# Samarati, P., & Sweeney, L. (1998). Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement Through Generalization and Suppression. Technical Report SRI-CSL-98-04, SRI International.

# pandas — pandas documentation // Documentation URL: <https://pandas.pydata.org/docs/> (дата обращения: 17.10.2024).

# tkinter — Python interface to Tcl/Tk // Documentation URL: <https://docs.python.org/3/library/tkinter.html> (дата обращения: 17.10.2024).